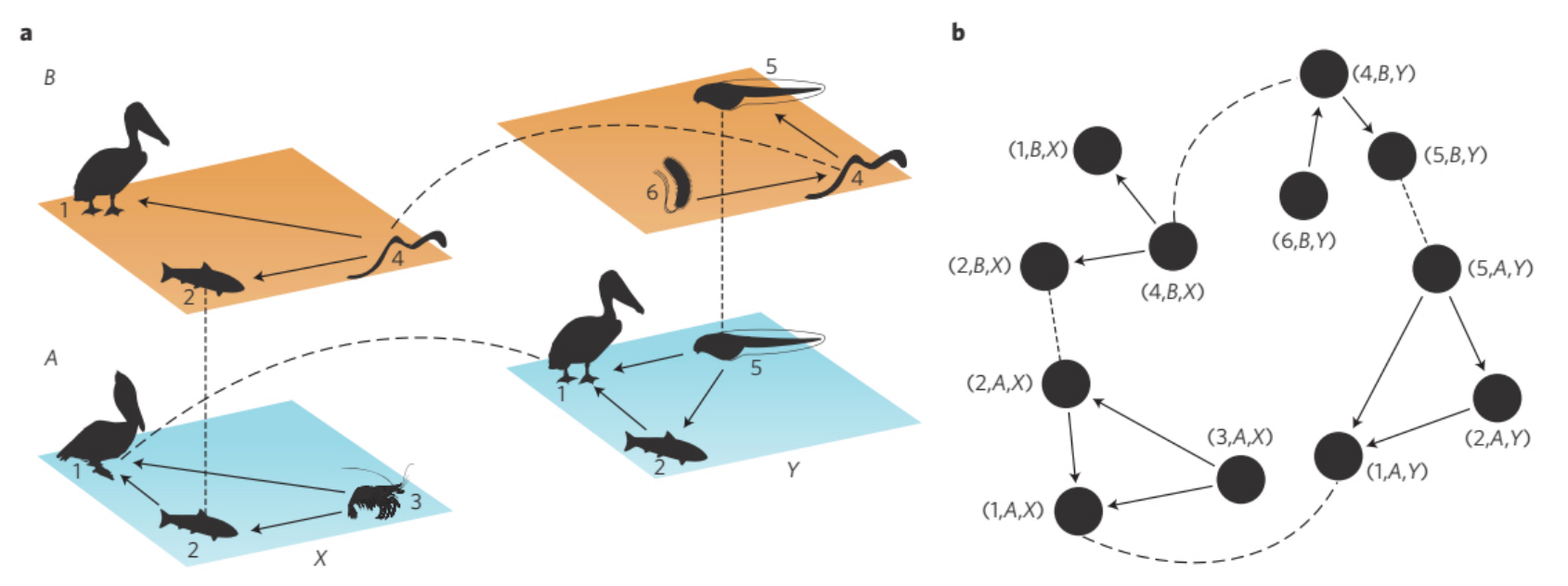
Tracking the data provenance for multilayer network modeling and analysis of foodwebs

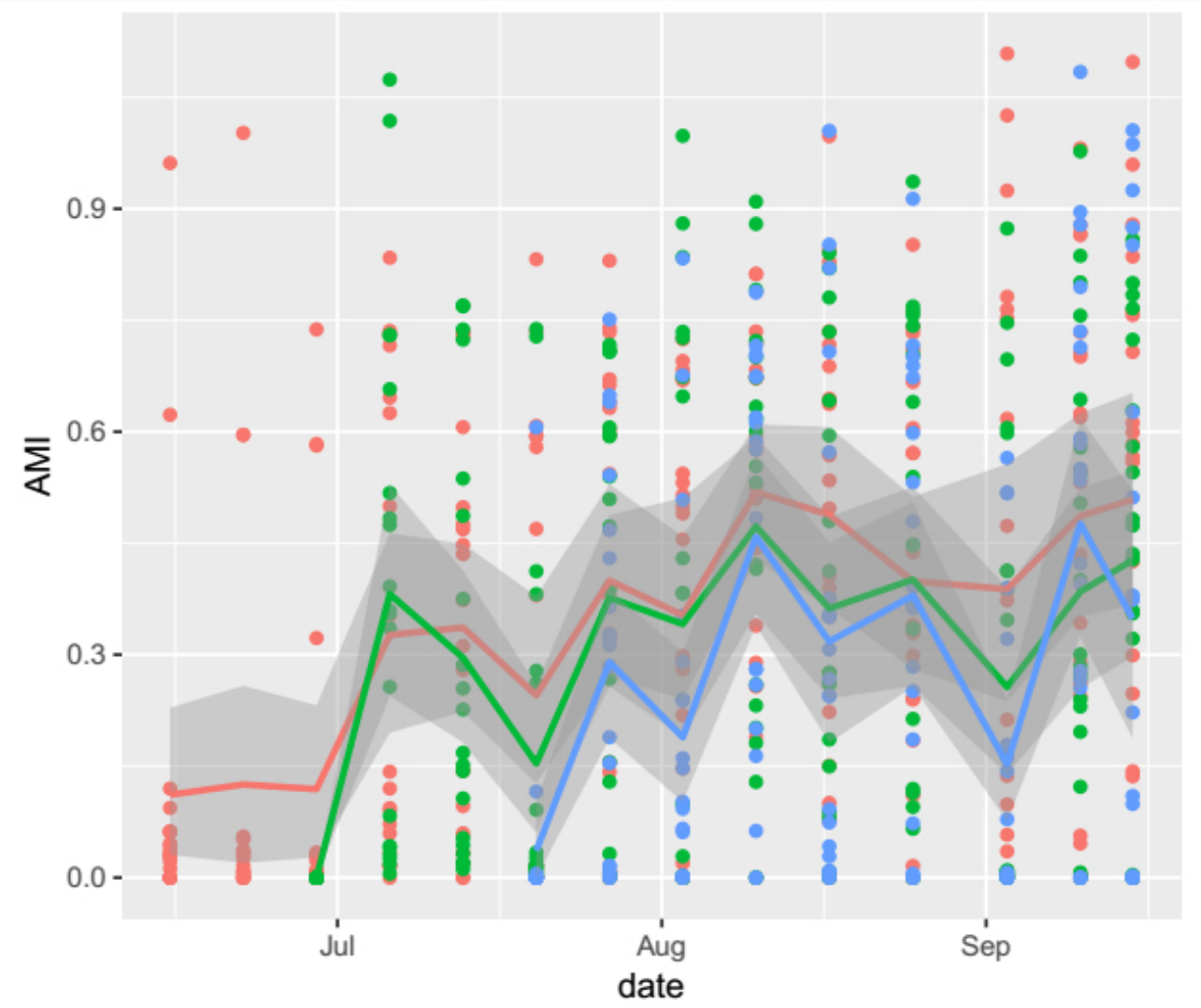
Matthew K. Lau, Thomas F.J.M. Pasquier, Aaron M. Ellison (Harvard University)

Pitcher Microecosystems

Multilayer network analysis



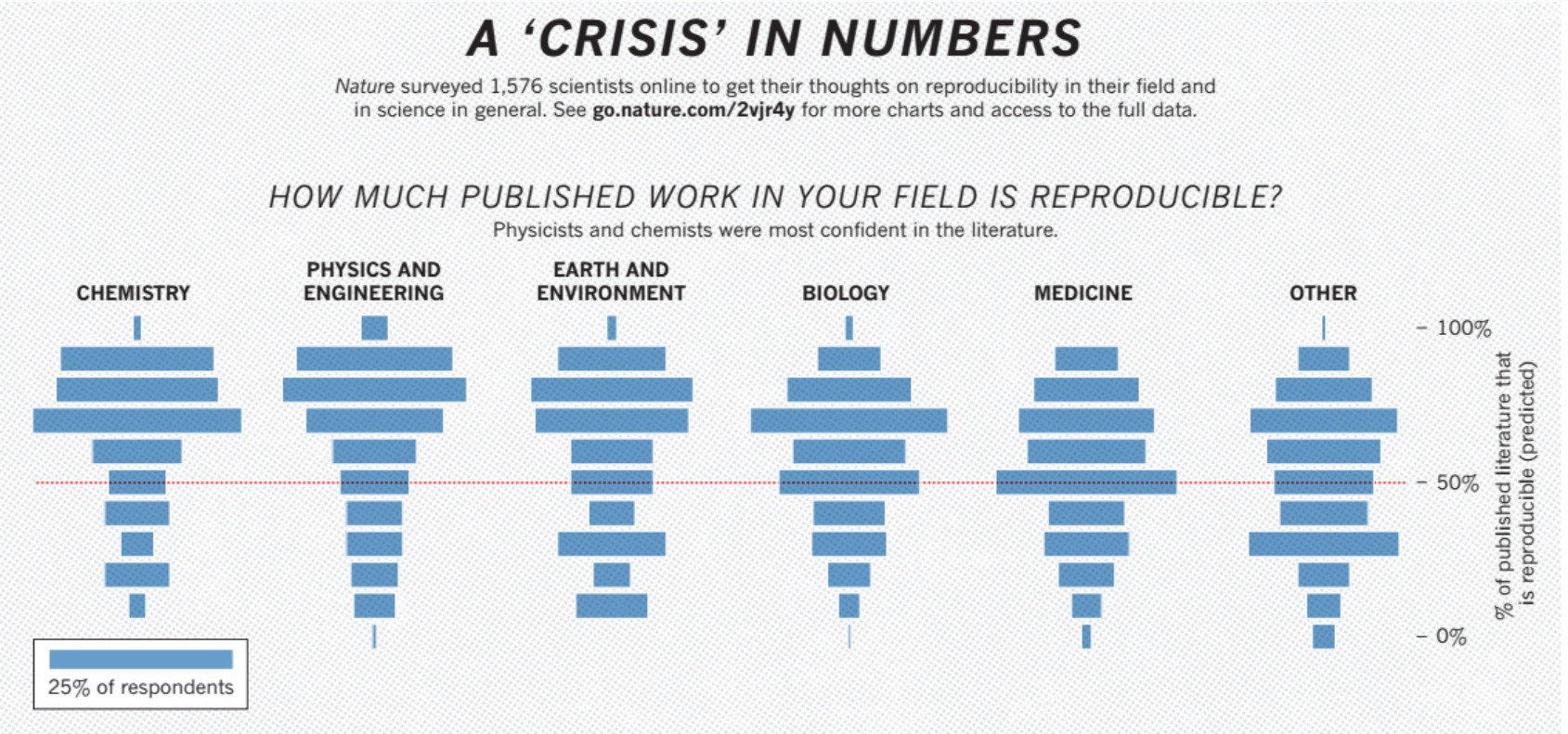
Multilayer networks



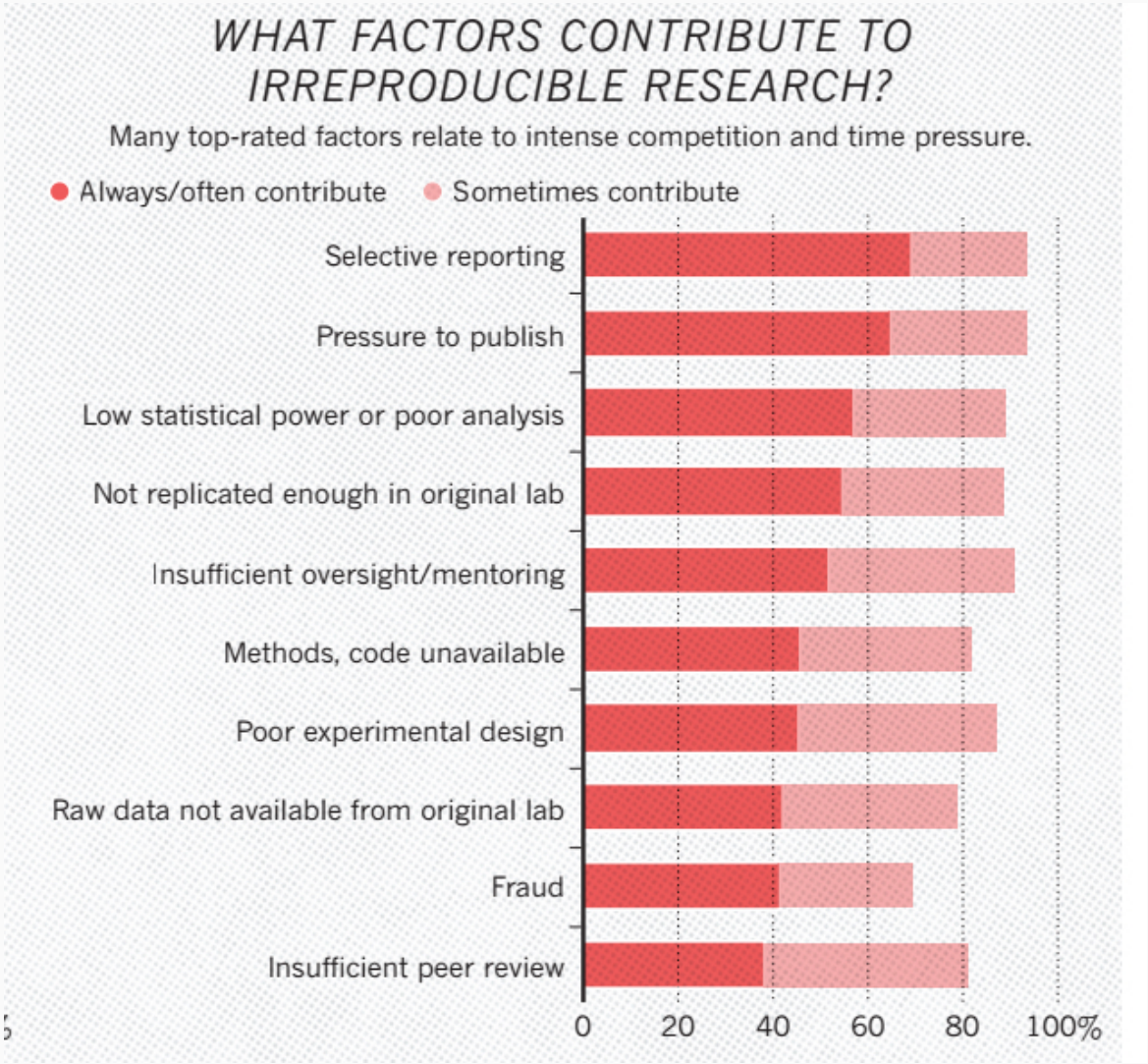
Data availability. The raw data for the example temporal network are in Supplementary Data 1 and deposited in Figshare (https://dx.doi.org/10.6084/m9.figshare.3472646.v2). The code for general procedures to prepare data, manipulate networks, post-process modularity-maximization calculations, and analyse network robustness are written in R, and they are available at https://dx.doi.org/10.6084/m9.figshare.3472664.v1. The code for examination of modular structure is written in MATLAB and available at https://dx.doi.org/10.6084/m9.figshare.3472679.v1.

Received 20 April 2016; accepted 27 January 2017;
published 23 March 2017

A Reproducibility Crisis (Baker 2015)



A Reproducibility Crisis (Baker 2015)



Open Process (Pasquier et al. 2017)

Open Process (Pasquier et al. 2017)

www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

Comment: If these data could talk

Thomas Pasquier[1], Matthew K. Lau[2], Ana Trisovic[3,4], Emery Boose[2], Ben Couturier[3], Mercè Crosas[5], Aaron M. Ellison[2], Valerie Gibson[4], Chris Jones[4] & Margo Seltzer[1]

Received: 12 April 2017
Accepted: 24 July 2017
Published: xx xxx 2017

In the last few decades, data-driven methods have come to dominate many fields of scientific inquiry. Open data and open-source software have enabled the rapid implementation of novel methods to manage and analyze the growing flood of data. However, it has become apparent that many scientific fields exhibit distressingly low rates of repeatability and reproducibility. Although there are many dimensions to this issue, we believe that there is a lack of formalism used when describing end-to-end published results, from the data source to the analysis to the final published results. Even when authors do their best to make their research and data accessible, this lack of formalism reduces the clarity and efficiency of reporting, which contributes to issues of reproducibility. Data provenance aids both repeatability and reproducibility through *systematic* and *formal* records of the relationships among data sources, processes, datasets, publications and researchers.

Challenge: How do you know what the code acutally does?

```
q <- runif(1)
if (q > 0.5){
     print("Yay!")
   }else{
     print("boooo!")
}
```

Data Provenance and Ecological Networks

Methods in Ecology and Evolution *Explore this journal >*

View issue TOC
Volume 5, Issue 11
November 2014
Pages 1206–1213

Application

enaR: An R package for Ecosystem Network Analysis

Stuart R. Borrett, Matthew K. Lau

First published:
27 October 2014 Full publication history

DOI:
10.1111/2041-210X.12282 View/save citation

Cited by (CrossRef):
20 articles Check for updates

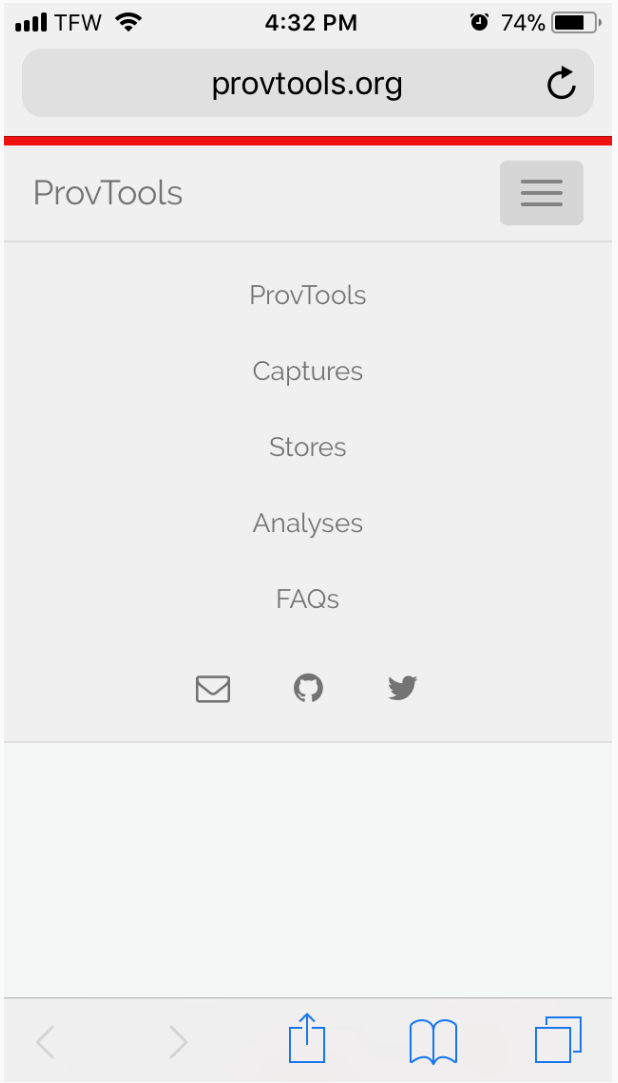
Citation tools

ProvTools

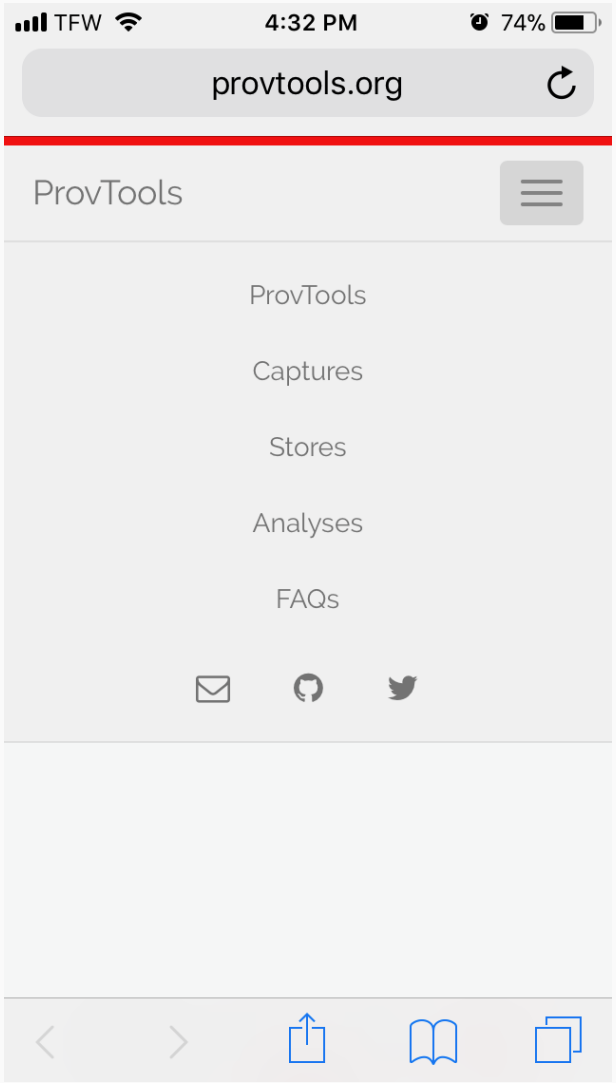
Capture and analyze provenance to improve data science

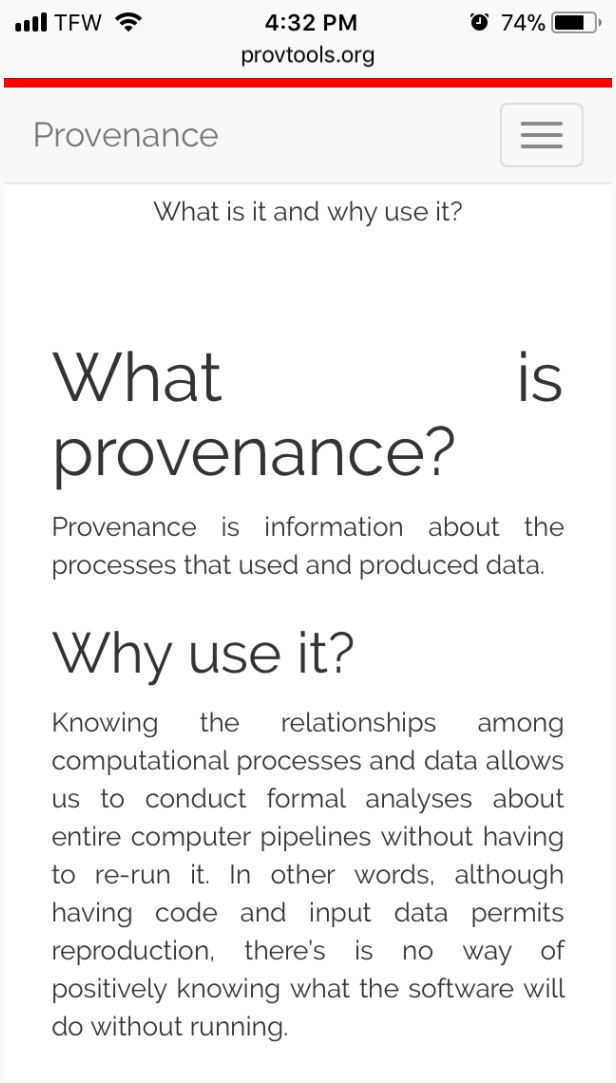
ProvTools


TFW
4:32 PM
74%
provtools.org
ProvTools
ProvTools
Capture and analyze provenance to improve data science


TFW
4:32 PM
74%
provtools.org
ProvTools
ProvTools
Captures
Stores
Analyses
FAQs

ProvTools


TFW
4:32 PM
74%
provtools.org
ProvTools
ProvTools
Captures
Stores
Analyses
FAQs


TFW
4:32 PM
74%
provtools.org
Provenance
What is it and why use it?
What is provenance?
Provenance is information about the processes that used and produced data.
Why use it?
Knowing the relationships among computational processes and data allows us to conduct formal analyses about entire computer pipelines without having to re-run it. In other words, although having code and input data permits reproduction, there's is no way of positively knowing what the software will do without running.

provR (aka. RDataTracker)

cleanR

```
18
19 ### Some datasets are loaded and no
       longer used.
20 ### Like this one
21 data.16.2 <- read.csv('projects/data_
       forestplot/dataset_v2_june_from_
       collaborator1.csv')
22
23 ### Create a bunch of intermediate
       objects
24 data.v1.1to4 <- data.16[,1:4]
25 data.v1.1to4. <- data.v1.1to4
26 data.v1.1to4 <- data.v1.1to4 * 2
27 data.v1.1to4.2 <- data.v1.1to4 * 2
28 data.16[,1:4] <- data.v1.1to4.2
29 library('vegan')
30 d1 <- vegdist(data.16[,1:2])
31 d2 <- vegdist(data.16[,2:3])
32
33 ### Conduct some analyses
34 mant1 <- mantel(d1,d2)
35 mant2 <- mantel(d2,d1)
36 mant11 <- mantel(d1,d1)
37 fit1 <- lm(Sepal.Length~Sepal.Width,data=
       data.16)
38 lm.summary.1 <- summary(fit1)
39
40 ### write some data to file
41 write.csv(data.v1.1to4,'projects/data_
       forestplot/save1.csv',row.names = F)
42
43 ### write some analyses to file
44 capture.output(lm.summary.1, file="
       analysis_forest/anova_table_1.txt")
45
46 ### write some figures to file
47 ### Here's another random, unused package
48 library('txtplot')
49
50 png('figures_1/fig1_biplot.png')
51 plot(data.16[,1:2])
52 dev.off()
53
54 png('figures_1/fig1_biplot_t2.png')
55 plot(data.16[,1:2]*2)
56 dev.off()
57
58 png('figures_2/fig2_biplot.png')
59 plot(data.16[,2:3])
60 dev.off()
```

Listing 7. Original "messy" code.

```
1 data.16 <- read.csv("projects/20
      biomass_survey.csv")
2 data.v1.1to4 <- data.16[, 1:4]
3 data.v1.1to4 <- data.v1.1to4 * 2
4 data.v1.1to4.2 <- data.v1.1to4 *
5 data.16[, 1:4] <- data.v1.1to4.2
6 png("figures_2/fig2_biplot.png")
7 plot(data.16[, 2:3])
8 dev.off()
```

Listing 9. Curated code for figure 2

encapsulator

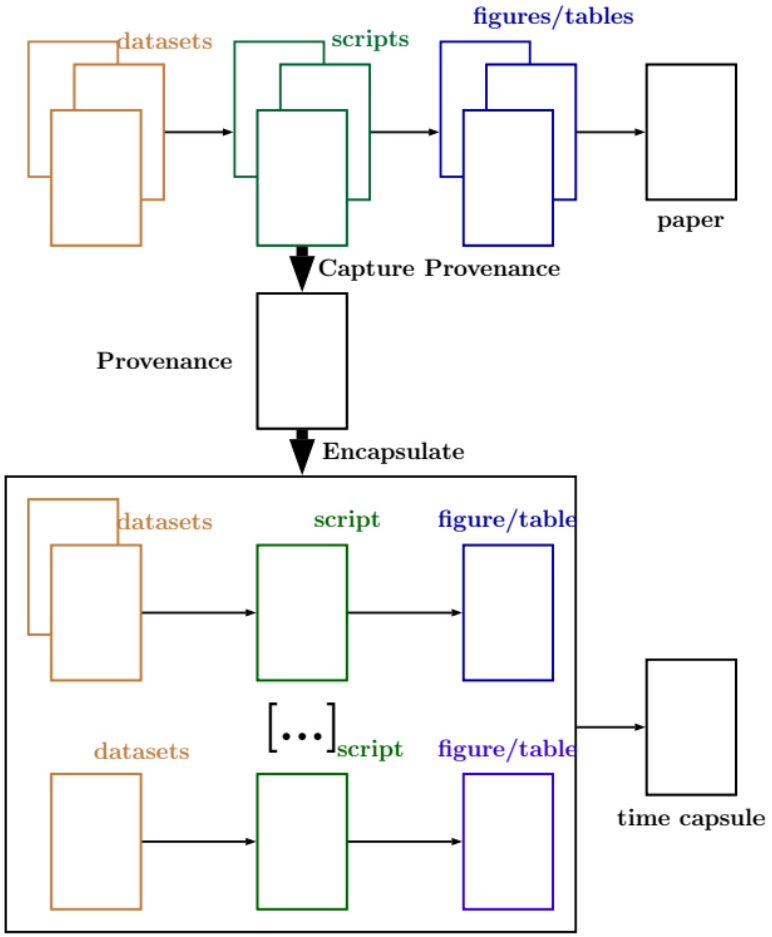


Figure 3. The encapsulation process.

enaR model packing (Borrett and Lau 2014)


```
library(enaR)   # load enaR package
> # -- ENTER MODEL DATA -- from Dame and Patten (1981)
> # node names
> names <- c("Filter Feeders","Microbiota","Meiofauna",
+                       "Deposit Feeders","Predators","Deposited Detritus")
> # Internal Flows of model, as matrix (oriented row to column)
> F <- matrix(c(0, 0, 0, 0, 0, 0, 0, 0, 0,
+               0, 0, 8.1721, 0, 1.2060, 0, 0, 0, 7.2745,
+               0, 1.2060, 0.6609, 0, 0, 0.6431, 0.5135, 0, 0,
+               0.1721, 0, 0, 15.7910, 0, 4.2403, 1.9076, 0.3262, 0),
+             ncol=6)
> rownames(F) <- names  # add node names to rows
> colnames(F) <- names  # add node names to cols
> # boundary flows
> inputs <- c(41.47,0, 0, 0, 0, 0)
> outputs <- c(25.1650, 5.76, 3.5794, 0.4303, 0.3594, 6.1759)
> # Living
> Living <- c(TRUE,TRUE,TRUE,TRUE,TRUE,FALSE)
> # pack the model data into the R network data object
> m <- pack(flow=F,input=inputs, respiration=outputs, outputs=outputs, living=Living)
>
> ssCheck(m)                # check to see if the model is at steady-state
[1] TRUE
> # perform flow analysis
> F <- enaFlow(m)               # perform ENA flow analysis
> attributes(F)                  # show analysis objects created
$names
[1] "T"  "G"  "GP" "N"  "NP" "ns"

> F$ns                          # show flow analysis network statistics
     Boundary     TST TSTp      APL       FCI       BFI       DFI       IFI
[1,]    41.47 83.5833   NA 2.015512 0.1101686 0.4961517 0.1950689 0.3087794
         ID.F   ID.F.I   ID.F.O    HMG.I    HMG.O AMP.I AMP.O mode0.F  mode1.F
[1,] 1.582925 1.716607 1.534181 2.051826 1.891638     3     1   41.47 32.90504
      mode2.F  mode3.F mode4.F
[1,] 9.208256 32.90504   41.47
> F$T
    Filter Feeders         Microbiota          Meiofauna    Deposit Feeders
           41.4700             8.1721             8.4805             2.5100
         Predators Deposited Detritus
            0.6856            22.2651
```

Policy: *OPEN* Science = data + source + process

Ocean & Coastal Management 115 (2015) 71–76

Contents lists available at ScienceDirect

Ocean & Coastal Management

journal homepage: www.elsevier.com/locate/ocecoaman

"Back off, man, I'm a scientist!" When marine conservation science meets policy

Naomi A. Rose [a], E.C.M. Parsons [b, *]

[a] Animal Welfare Institute, Washington DC, USA
[b] Department of Environmental Science & Policy, George Mason University, Fairfax, VA, USA

ARTICLE INFO

Article history:
Received 4 January 2015
Received in revised form
24 April 2015
Accepted 27 April 2015

ABSTRACT

There is often a basic tension at the boundary between science and policy – the former seeks unbiased, objective descriptions of reality, while the latter must incorporate various factors in its development, including values, ideologies, economics, biases, and emotions. Problems may arise if, and when, marine scientists who enter the policy arena fail to understand these differing priorities, and we describe some

Policy: *OPEN* Science = data + source + process

Troublesome scientists

Here are three 'types' of scientists who can create difficulties (rather than assist in solutions) when involving themselves in policy debates and discussions.

The Naïve Scientist

These scientists believe that if only policy-makers had the right information, they would make the right decisions. They do not understand the important human and legal dimensions of policy-making and make little attempt to interpret their work in that context. As one researcher stated in an interview, "*If all sides devoted their resources to research rather than to lawsuits, we could get some answers, but without them, the lawsuits will continue*" (Madin, 2009).

The 'Ivory Tower' Scientist

They believe that it is essential for scientists to remain 'pure', to stay at arm's length from anything resembling advocacy, even though they may also seek media attention for their work. This may result in essential data not reaching policy-makers, resulting in poor decisions. Or, even worse, their research is mischaracterized or misunderstood and they make no attempt to correct these misinterpretations (see below). As a researcher once told one of the authors (Rose): "*I can't be held responsible for the policy implications of my work.*"

The 'Industry' Scientist

They work directly for special interests, but expect their science (even when not peer-reviewed) to be accepted as objective. They either do not understand their conflicts of interest, or ignore them. As one government scientist emphatically stated in a policy meeting attended by one of the authors (Rose), in response to a comment from a participant that there was disagreement over industry research results, "*Only if you disagree with science!*"

Thank you

Interested in alpha testing or more information about provenance and ProvTools:

provtools.org

Email me: *matthewklau@fas.harvard.edu*